arXiv:2611.32938v1 [cs.CV] 24 Nov 2026

通过令牌分配实现高效视频检索增强生成

Ayush Barik[1] Fan Lai[1†] Francis Y.Yan[1†]伊利诺伊大
学厄巴纳-香槟分校{barik2, fanlai, fyy}@illinois.edu

摘要

高效的长视频理解需要识别相关证据，同时尽量减少视觉语言模型（*VLM*）处理的词元数量。现有的视频检索增强生成（*RAG*）方法依赖帧选择和词元压缩，却忽视了每增加一个词元所带来的价值。我们提出***Lembas***，这是一个即插即用框架，将视频检索增强生成建模为在共享预算下对帧和字幕进行词元分配的问题。我们的核心洞见是，词元在每个项目中的价值会逐渐递减；当相似证据获得更多词元时，其价值还会进一步降低。*Lembas*通过一个目标函数同时刻画这两种效应，并采用高效的迭代注水算法求解，且该算法具有可证明的收敛保证。对于多轮对话，*Lembas*会复用先前轮次已处理的词元，并利用视觉语言模型的注意力来优化相关性，无需额外成本。评估结果表明，在计算量相当的情况下，*Lembas*的准确率比当前最先进基线提高最多*7.2*个百分点。在多轮视频理解任务中，*Lembas*的准确率最高提升*1.5*个百分点，同时减少*1.7–17.3*× 万亿次浮点运算。

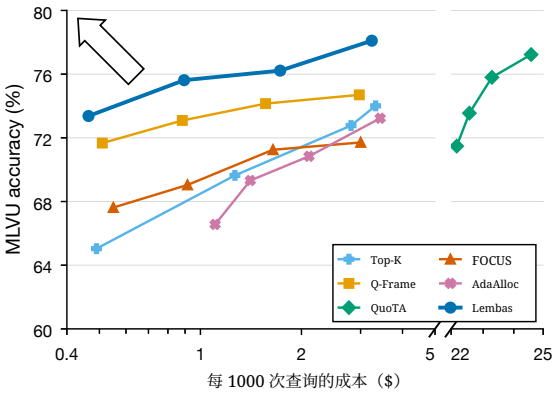

图 1。Lembas 实现了最佳的准确率—成本权衡。美元成本根据 OpenRouter API 对 Qwen3-VL-8B 的定价计算得出。

1. 引言

视觉语言模型（VLMs）在视频理解方面取得了显著进展，能够执行从动作识别到复杂视觉问答等任务 [2, 47]。然而，将这些能力扩展到长视频（如体育赛事转播、个人录像和监控录像）仍然面临挑战。处理每一帧很容易超出视觉语言模型的上下文窗口，并产生高昂的推理成本 [6, 56]。高效的长视频理解需要识别并处理对回答每个查询最有用的证据。

检索增强生成（RAG）通过检索相关帧而非处理整个视频来应对这一挑战。基于嵌入的检索增强生成会检索与查询的语义相似度较高的帧，而基于图的检索增强生成则将实体和事件连接起来进行检索
相关证据 [32, 38, 41, 55]。然而，这些优化方法虽然能够选择相关的证据项目，却常常忽视了一个固有区别：不同查询所需的证据量和视觉细节各不相同。例如，识别一个小物体可能需要来自特定帧的细粒度信息，而重建一系列事件则可能受益于更广泛的时间覆盖范围。

近期方法通过自适应帧选择和令牌压缩来改善这一权衡。帧选择方法利用检索启发式方法
[37, 42, 44, 49, 57, 65], 、学习到的策略 [5, 27, 58, 59], 或智能体搜索 [48, 66]来识别相关帧或具有多样性的帧。令牌压缩方法通过令牌合并或剪枝 [11, 40, 45]来减少视觉冗余。更新的方法则通过按相关性比例分配 [33]、离散分辨率层级 [62], 或智能体预算规划 [1]来调整令牌数量。尽管取得了这些进展，分配仍主要由独立的相关性分数或预定义规则指导。

我们的关键洞见是，令牌效率要求根据令牌的边际价值进行全局分配。我们有两点观察。首先，额外令牌呈现边际收益递减：一旦某帧保留了足够的细节，继续分配更多令牌所带来的准确性提升会逐渐减小——

†共同承担指导工作。



图 2. 调整每帧令牌数，使 Lembas 能够在共享预算内保留更多帧中的证据。

益。其次，令牌价值取决于其他已分配的证据：当相似帧已经获得令牌时，某一帧贡献的新信息会减少。自然地，一个根本问题是如何在帧和字幕之间分配共享的令牌预算，以最大化它们的总体效用。

基于这些观察，我们提出 Lembas，这是一种用于令牌高效视频检索增强生成的即插即用框架（图 3）。给定现有检索增强生成系统检索到的帧和字幕，Lembas 会评估它们的效用、分配共享令牌预算，并在视觉语言模型推理前相应地剪枝每个项目。实现这一设计面临两项挑战：如何建模令牌价值，以及如何高效优化分配方案。此外，用户很少只针对一段视频提出一个问题，而跨越多轮的对话又带来了第三项挑战：随着各轮分配方案的变化，如何重复利用证据。

令牌价值建模。某个项目的贡献不仅取决于它与查询的相关性，还取决于它获得的令牌数量，以及其内容有多少已被其他项目覆盖。我们考察下游准确率随令牌数量的变化，发现凹饱和有理函数能够很好地刻画边际收益递减（图4）。我们将这种建模与查询相关性和冗余感知相结合，并使用基于注意力的剪枝来实现由此得到的令牌数量（§3.2）。

令牌分配。考虑冗余会使令牌分配彼此依赖，从而导致组合空间呈爆炸式增长。我们提出迭代注水法，反复更新多样性感知相关性，并求解一个闭式注水子问题。我们的求解器收敛于连续松弛的全局最优解，其 $O(1/\ell)$ 目标函数差距经过 ℓ次迭代后达到3.3节）。

跨轮次复用证据。多轮对话经常会重新访问相同的证据，但不同查询可能需要不同的令牌数量，从而使 KV缓存复用变得复杂。Lembas 通过查询无关令牌排序解决了这一问题，使后续轮次只需预填充缓存中缺失的词元。此外，我们利用视觉语言模型对缓存证据的注意力，在无需额外视觉语言模型调用的情况下，优化其在后续分配中的相关性（§3.4）。

我们在 MLVU、LongVideoBench 和 Video-MME 上，使用 Qwen3-VL-8B 和 InternVL3.5-4B 对 Lembas 进行评估（§4）。在计算量预算相当的情况下（图 1 和表 1），与当前最先进基线相比，Lembas 将准确率提高了 0.1–7.2 个百分点。在多轮基准测试 MT-Video-Bench 上，与基线方法相比，Lembas 的准确率高出 0.3–1.5 个百分点，同时减少了 1.7–17万亿次浮点运算。

总体而言，本文的贡献如下：

•我们提出了 Lembas，这是一种令牌高效视频检索增强生成框架，将视频理解建模为预算有限条件下的令牌分配问题。

• 我们开发了令牌分配器，利用迭代式注水法高效求解令牌优化问题。

• 我们通过增量分配和注意力细化，将 Lembas 扩展至多轮对话，使系统能够在分配方案变化时复用证据。

2. 相关工作

面向视频理解的检索增强生成。视频检索增强生成 [28] 将检索增强生成扩展到长视频，而长视频的帧数远超视觉语言模型的上下文窗口。现有系统的主要区别在于如何为视频建立索引，以及如何从索引中检索证据。在索引方面，一类方法将视频转换为文本代理，例如字幕、语音转录文本和屏幕文字，并在这些内容上进行检索 [32, 34], ；另一类方法则直接对帧或片段进行嵌入，并根据查询与视觉内容的相似性进行检索，包括跨视频语料库 [23]和直播流 [54]的检索。为捕捉此类扁平索引所遗漏的关系，结构化索引还会将视频组织成树 [50] 或实体与事件中心图 [38, 41, 55]。在检索方面，单步相似性搜索已扩展为多步检索和智能体检索，后者通过多轮迭代不断优化检索到的证据 [41, 55]。然而，这些系统改进的是索引和检索的内容，并以统一且完整的词元成本将每个检索到的项目传递给视觉语言模型。Lembas 则关注每个检索到的项目应分配多少词元，并作为即插即用模块，在这些检索器的下游运行。

关键帧或令牌选择。与其完整处理长视频来回答用户问题，各种方法会智能地决定处理哪些帧 [29, 30, 37, 42–

44, 49, 59, 65], 这可能是由于实际约束（例如上下文窗口 [61]），也可能是为了通过降低令牌预填充成本来提高效率 [6, 56]。令牌剪枝已在长上下文大语言模型服务中得到广泛研究。H2O [63]会剪去累积注意力较小的令牌，而 StreamingLLM [52] 则保留初始令牌和最近的令牌。除了决定保留哪些帧或令牌之外，近期的相关工作还会决定在每个帧上投入多少，通常通过智能体搜索 [1] 或手工设计的规则 [33, 62]将帧路由到少数固定的分辨率层级。Lembas 则将这一问题视为资源分配问题，构建目标函数，在具有收敛保证的情况下求解，并将其扩展到多轮对话。

传统资源分配。 资源分配已在网络和计算系统中得到广泛研究，用于在相互竞争的需求之间分配有限资源 [13, 26, 31]。经典方法包括效用最大化和定价机制 [7, 35], 、注水法和功率控制 [14, 39, 60], , 以及公平性感知调度 [16, 18, 21]。Lembas将这种资源分配视角引入视频理解，通过迭代注水，在检索到的项目之间分配有限的令牌资源。

3. 方法

Lembas 是一个即插即用框架，可提高视频检索增强生成的令牌效率。如图 3所示，给定查询后，Lembas 会评估每帧和字幕的相关性，在令牌预算内优化它们的分配，并从每个选定项目中保留排名最高的令牌。其令牌分配器利用令牌价值的边际收益递减特性，并降低冗余证据的权重。对于多轮对话，Lembas 还会复用 KV 缓存中已有的令牌，并利用视觉语言模型的注意力来调整其相关性。

3.1. 问题表述

给定视频 V和查询 q。传统的 Top-K 视频检索增强生成会检索与 q相关性最高的 K 帧。令$f_1, \dots, f_N$为候选帧，并令 $r_i(q) \geq 0$ 表示帧 i的检索相关性。从优化的角度来看，这是一个整数线性规划：

$$\max_{\mathbf{x}\in\{0,1\}^N} \sum_{i=1}^{N} r_i(q) x_i \quad \text{s.t.} \quad \sum_{i=1}^{N} x_i \leq K, \tag{1}$$

其中， x_i 表示是否选中帧 i 。最优解只需选取评分最高的 K 帧。然而，每个选中的帧都按固定的令牌数量进行处理。因此，Top-K 无法舍弃某一帧中不必要的细节，以换取另一帧中的有用证据。

Lembas 旨在打破这种僵化，将问题从帧选择提升至多模态令牌分配，从而提高效率。

设 $\mathcal{F}$ 和 $\mathcal{S}$ 分别表示候选帧和字幕片段， $b_i \in [0, b_i^{\max}]$ 表示分配给项目 $i \in \mathcal{F} \cup \mathcal{S}$ 的词元，其中 $b_i^{\max}$ 是其完整令牌数量。函数 $v_i(b_i) \in [0,1]$（对帧而言为 v_{frame}，对字幕而言为 v_{text}）用于建模项目 i 以 b_i 个词元表示时保留的价值；而 $r_i(q,\mathbf{b})$ 则衡量在考虑查询及分配至其他位置的证据后，该项目的相关性。帧和字幕共享总预算 B，其中 $\beta > 0$ 用于平衡二者的贡献。随后，Lembas 旨在优化：

$$\begin{aligned} \max_{\mathbf{b}} \quad & \sum_{i\in\mathcal{F}} r_i(q,\mathbf{b})\, v_{\text{frame}}(b_i) + \beta \sum_{i\in\mathcal{S}} r_i(q,\mathbf{b})\, v_{\text{text}}(b_i) \\ \text{s.t.} \quad & \sum_{i\in\mathcal{F}\cup\mathcal{S}} b_i \leq B \end{aligned} \tag{2}$$

然而，这种表述在实践中面临三项挑战： 如何对令牌价值函数 v_{frame}和 v_{text} 以及每个项目的相关性进行建模 r_i？（§3.2） 如何高效地优化所述目标函数？（§3.3）以及如何在多轮对话中优化并复用已分配的令牌？（§3.4）。

3.2. 构建分配目标函数

令牌价值函数。 检索相关性估计项目在完整表示时的相关性，但无法体现令牌剪枝后其有效性有多少得以保留。在公式（2）中，我们使用令牌价值函数 v_{frame}（b_i） 和 v_{text}（b_i）进行建模：这些函数根据令牌数量 b_i时保留的价值比例，对项目的相关性进行加权。为实现任意令牌数量 b_i，Lembas 对每个帧的视觉令牌进行一次离线排序，该排序与查询无关，依据的是它们在视觉语言模型视觉编码器最后一层中获得的注意力，并对所有头进行求和；随后按照流行令牌剪枝器所采用的基于注意力的排序方法，保留排名最高的 b_i个令牌，并保持其原始位置 [6, 56]。

为了定义令牌价值函数，我们测量问答准确率随令牌数量的变化。在多个视频问答基准的留出验证集上 [10, 51, 64], , 我们将一段短至中等长度视频的每个帧剪枝至 b_i 个令牌，将这些帧和查询输入视觉语言模型，并记录每个 b_i处的平均准确率。随后，我们对准确率进行归一化，并拟合若干候选函数（图 4）。对于视觉和文本证据，一个简单的有理函数最符合测量结果：

$$v_{\text{frame}}(b_i) = \frac{b_i}{b_i + \tau_{\text{frame}}}, \quad v_{\text{text}}(b_i) = \frac{b_i}{b_i + \tau_{\text{text}}}, \tag{3}$$

其中， $\tau_{\text{frame}} > 0$ 和 $\tau_{\text{text}} > 0$ 是通过最小二乘法离线估计的特定于模态的参数。拟合结果达到$R^2 > 0.85$，这表明，对于此任务而言，简单函数就已足够，无需使用更复杂的方法（例如神经网络）。由于静态函数可能无法适应查询分布的变化，因此可以在

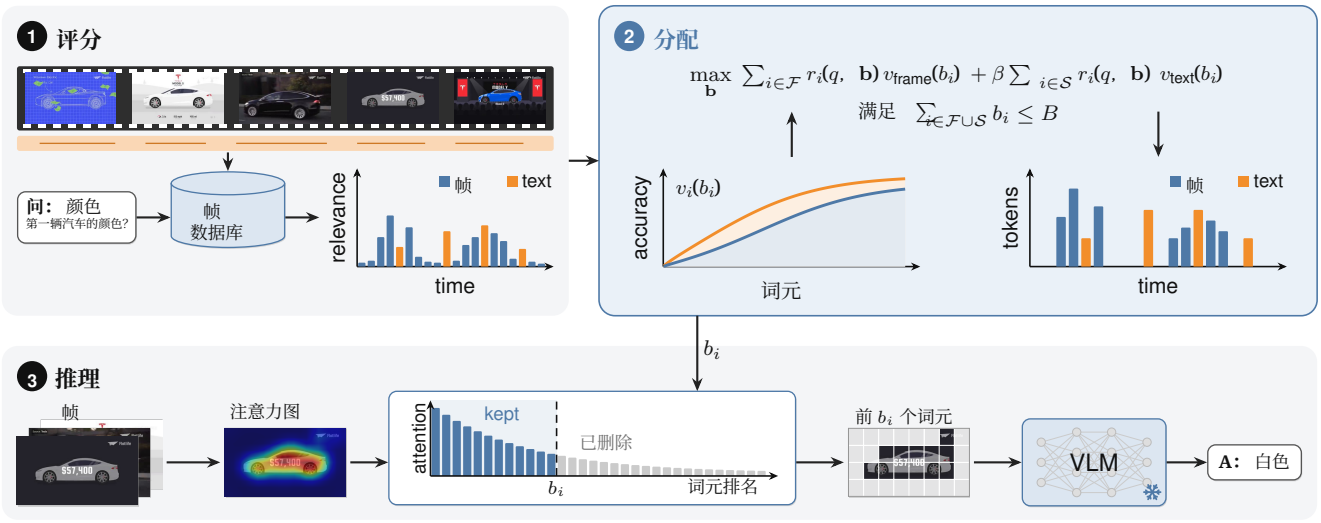

图 3。Lembas 工作流程示意图。

查询反馈（例如准确率）可用的非高峰服务时段重新拟合这些参数。

多样性感知相关性。 检索相关性$r_i(q)$能够捕捉相关性，但假设项目彼此独立。然而，在视频中，相邻帧可能呈现几乎相同的内容；因此，应降低与已选定帧相似的帧的权重。为捕捉这种依赖关系，我们根据已分配给相似项目的令牌价值，对每个项目的检索相关性$r_i(q)$进行折减。令 s_{ij} 表示同一模态内项目嵌入之间的余弦相似度。我们将完整的、依赖于分配的相关性定义如下：

$$
\begin{aligned}
r_i(q,\mathbf{b}) &= r_i(q) - \eta \sum_{j\in\mathcal{F}} s_{ij}\, v_{\text{frame}}(b_j), \quad i \in \mathcal{F}, \\
r_i(q,\mathbf{b}) &= r_i(q) - \eta \sum_{j\in\mathcal{S}} s_{ij}\, v_{\text{text}}(b_j), \quad i \in \mathcal{S},
\end{aligned}
\tag{4}
$$

其中，$\eta \ge 0$ 控制冗余惩罚。

将式（3）和（4）代入式（2），最终得到完整的分配目标函数。它会奖励与查询相关的证据，考虑令牌价值递减，并抑制冗余分配。

3.3. 高效令牌分配

直接分配离散的令牌数量涉及许多整数决策，而这些决策通过多样性项相互耦合。现成的整数求解器（例如 Gurobi [17],CPLEX [20]）在最坏情况下需要指数级时间：对于一段一小时的视频样本，在 1 帧/秒下，它们需要运行十几秒以上，超过了视觉语言模型本身的推理时间。

因此，我们改为优化连续松弛，然后将其解舍入为完整令牌。关键在于

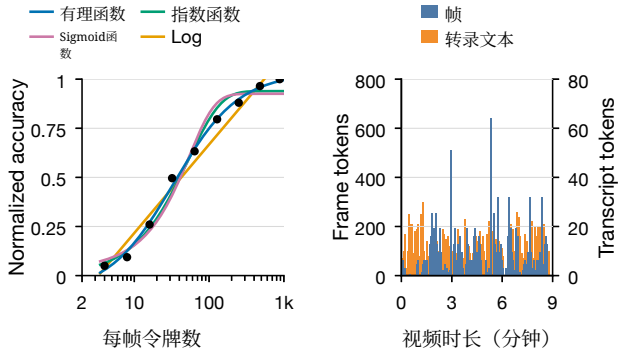

图 4. 候选令牌价值函数，拟合于测得的帧准确率（黑点）。图 5. Lemma 的令牌分配示例，针对一段九分钟的视频。

将变量从令牌数量改为令牌价值$z_i = v_i(b_i)$。由于我们的有理令牌价值函数是凹函数，且相似度矩阵是半正定的，这种松弛将式（2）转化为关于令牌价值的连续凸规划。首先，我们在相关性固定的情况下求解一个更简单的子问题，这自然引出了注水法算法；随后，我们取消这一限制，并将其扩展为针对完整目标函数的迭代注水法。

注水法。 首先考虑这样一个子问题：每个项目的相关性固定为 $\bar{r}_i$，其检索相关性为 $r_i(q)$（字幕按 β 进行缩放），因此不再取决于分配。此时，目标函数可化简为在单一预算约束下对凹项 $\bar{r}_i\, v_i(b_i)$ 求和，其中为某个项目分配的每个额外令牌所带来的增益都小于前一个令牌。因此，在最优值处，所有获得部分分配的

项目必须具有相同的每令牌边际增益，我们将其记为 λ；否则，将令牌从边际增益较低的项目移至边际增益较高的项目，就能改善目标函数。对于我们的令牌价值函数，项目 i 的边际增益为 $\bar{r}_i\tau_i/(b_i+\tau_i)^2$,，其中 τ_i 是其模态的令牌价值参数。令其等于 λ 并求解 b_i，可得

$$b_i = \left[\sqrt{\bar{r}_i\tau_i/\lambda} - \tau_i\right]_0^{b_i^{\max}}, \tag{5}$$

其中 $[\cdot]_0^{b_i^{\max}}$ 将令牌数量截断至 $[0, b_i^{\max}]$。直观而言，只有当某个项目的相关性超过 $\lambda\tau_i$ 时，它才会获得令牌；此后，其令牌数量会随相关性的平方根增长，而非与相关性成正比。因此，单个标量 λ便同时决定了选取哪些项目，以及每个项目获得多少令牌。由于总分配量随 λ 单调递减，因此可以通过二分法找到满足预算 B的 λ [4, 60]。这一过程称为注水法。

迭代注水法。然而，完整目标函数会根据每个项目相似项目的令牌价值，对该项目的相关性进行折扣，因此注水法所需的相关性取决于其自身产生的分配结果。我们使用 Frank–Wolfe 方法 [9, 22], 解决这种循环依赖，该方法通过迭代优化目标函数的线性近似来实现。在令牌价值方面，式（2）可写为 $J(\mathbf{z}) = \bar{\mathbf{r}}^\top\mathbf{z} - \eta\mathbf{z}^\top\mathbf{S}\mathbf{z}$，其中 $\mathbf{S}$ 是模态内相似度的块对角矩阵 s_{ij}（字幕的相似度按 β 缩放）。在第 ℓ次迭代中，给定当前令牌价值 $\bar{\mathbf{z}}$（初始化为零），我们计算多样性折扣后的相关性如下：

$$\widetilde{\mathbf{r}} = \nabla J(\bar{\mathbf{z}}) = \bar{\mathbf{r}} - 2\eta\,\mathbf{S}\bar{\mathbf{z}}. \tag{6}$$

在令牌预算约束下最大化线性化目标函数 $\bar{\mathbf{r}}^\top\mathbf{z}$，恰好就是上述固定相关性子问题。因此，以 $\bar{r}_i$ 替代 $\widetilde{r}_i$ 进行一次注水迭代，即可得到其最优分配$\mathbf{b}^\star$；其中，满足 $\widetilde{r}_i \le 0$ 的项目不会获得令牌。然而，直接采用这一分配可能会过度修正冗余问题，因为它仅优化线性化目标函数。因此，我们采用递减步长，将令牌价值更新为$\mathbf{z}^\star = \mathbf{v}(\mathbf{b}^\star)$:

$$\bar{\mathbf{z}} \leftarrow (1-\gamma_\ell)\bar{\mathbf{z}} + \gamma_\ell\mathbf{z}^\star, \qquad \gamma_\ell = \frac{2}{\ell+1}. \tag{7}$$

第一次迭代会恢复标准注水法，而后续迭代则会进一步优化分配方案，以考虑冗余证据。在经过 L 次迭代后，我们得到令牌数量： $b_i = v_i^{-1}(\bar{z}_i)$。完整流程总结于算法1中。接下来我们将证明，我们的解以次线性速率实现全局收敛：

命题1。 设 $J^\star$ 是方程(2)在连续令牌分配上的全局最大值，并设 $J^{(\ell)}$ 是算法 ℓ 经过 ℓ 次迭代后的目标值，其中算法为*1*。则

$$J^\star - J^{(\ell)} \le \frac{4\,\eta\,\sigma_{\max}D^2}{\ell+2} = O\left(\frac{1}{\ell}\right), \tag{8}$$

其中， $\sigma_{\max}$ 是帧相似度矩阵与 β-缩放字幕相似度矩阵的最大特征值， D 是预算为 B时可行令牌价值集合的直径。

该界来自标准的 Frank–Wolfe 保证 [22]。视觉语言模型使用完整令牌，因此我们将连续分配向下取整， $b_i = \lfloor b_i \rfloor$。取整后的分配仍在预算范围内，并且每个项目移除的令牌数少于一个。最后，Lembas 为每个项目保留预先排序的前 b_i个令牌，并按时间顺序呈现所选证据。

相较于先前分配规则的最优性。 先前的方法按照固定规则分配令牌，而不是通过优化目标函数进行分配。QuoTA [33]为每个帧分配与其相关性成比例的令牌数量，而 Q-Frame [62] 则为相关性层级内的所有帧分配相同的令牌数量。即使忽略多样性，这两种规则都不是最优的。由- 于令牌价值函数是严格凹的，公式（5）中的注水分配是固定相关性目标的唯一最大化解，并且不同于上述两种规则。具体而言，其令牌数量随相关性的平方根增长，而不是与相关性成比例增长；此外，具有$\bar{r}_i \le \lambda\tau_i$ 的项目完全不会获得令牌。相比之下，比例分配会在最相关帧的价值达到饱和后仍继续向其添加令牌，而分层分配则忽略同一层级内相关性的差异。此外，这两种规则都会独立地评估帧，而迭代式注水法还会考虑项目之间的冗余，并收敛到完整目标函数的最优值（命题 1）。

3.4. 扩展至多轮对话

在实际对话中，用户可能会针对同一段视频提出多个查询q_1、.....、 q_T。 [19]后续查询通常会引用先前轮次处理过的证据（图6）。如果选择完全匹配，Top-K 可以复用先前预填充的帧，但令牌分配更为复杂：同一项目在不同轮次可能获得不同的令牌数量。Lembas 通过增量分配解决这一问题：它只预填充当前分配与所有先前分配之间的差额，并通过注意力细化免费利用注意力重新细化缓存证据的相关性。

算法 1迭代注水

输入： 相关性 $\bar{r}_i$、相似度 s_{ij}、价值参数 τ_i、令牌限制 $b_i^{\max}$、预算 B、惩罚项 η、迭代次数 L、容差 ϵ
输出： 令牌数量 b_i 使用 $\sum_i b_i \le B$
索引为 i 的行适用于每个项目 $i \in \mathcal{F} \cup \mathcal{S}$。

初始化令牌价值 $\bar{z}_i \leftarrow 0$
对于 $\ell = 1, \ldots, L$ **执行**
  $\widetilde{r}_i \leftarrow \bar{r}_i - 2\eta \sum_j s_{ij}\, \bar{z}_j$ ▷ 折扣相关性，公式（6）
  $\lambda_{\text{lo}} \leftarrow 0$，$\lambda_{\text{hi}} \leftarrow$ 最大值$_i\, \widetilde{r}_i/\tau_i$ ▷ λ的搜索范围
  $b_i^\star \leftarrow 0$ ▷ 目前预算内的最佳分配
  当 $\lambda_{\text{hi}} - \lambda_{\text{lo}} > \epsilon$ **时** ▷ 通过二分法进行注水
    $\lambda \leftarrow (\lambda_{\text{lo}} + \lambda_{\text{hi}})/2$
    $b_i \leftarrow \left[\sqrt{\text{最大值}\ (0,\ \widetilde{r}_i)\ \tau_i/\lambda} - \tau_i\right]_0^{b_i^{\max}}$ ▷ Eq. (5)
    如果 $\sum_i b_i > B$ **则**
      $\lambda_{\text{lo}} \leftarrow \lambda$ ▷超出预算：提高 λ
    else
      $\lambda_{\text{hi}} \leftarrow \lambda$，$b_i^\star \leftarrow b_i$ ▷在预算范围内：较低的 λ
  $\gamma_\ell \leftarrow 2/(\ell+1)$ ▷ 步长
  $\bar{z}_i \leftarrow (1-\gamma_\ell)\ \bar{z}_i + \gamma_\ell\, v_i(b_i^\star)$ ▷ 平均，公式（7）
$b_i \leftarrow v_i^{-1}(\bar{z}_i) = \tau_i \bar{z}_i/(1-\bar{z}_i)$ ▷ 连续令牌计数
返回 $b_i \leftarrow \lfloor b_i \rfloor$ ▷ 向下取整到完整令牌

增量分配。 令 $b_i^{(t)}$ 表示第 t轮中项目 i 所需的令牌数量，令 $c_i^{(t-1)}$ 表示其最高排名的令牌中已在 KV缓存中的数量，其中$c_i^{(0)} = 0$。需要预填充的新令牌数量以及更新后的缓存状态为

$$p_i^{(t)} = \max\left(0, b_i^{(t)} - c_i^{(t-1)}\right), \qquad c_i^{(t)} = \max\left(c_i^{(t-1)}, b_i^{(t)}\right). \tag{9}$$

由于标准服务系统通常会在多轮对话之间保留 KV缓存，因此，较早轮次预填充的证据仍可供后续查询利用，而预算 B 适用于每轮分配的证据。由于令牌排序与查询无关，在第 t 轮分配了更多令牌的项目，只需新增纳入排名 $c_i^{(t-1)}+1, \ldots, b_i^{(t)}$ 的令牌。Lembas 会将这些令牌追加到现有上下文中，而不是插入较早的位置，否则会使后续 KV缓存条目失效。模型会保留适用于各项目的视觉或时间位置信息。如图7所示，增量分配因此能够恢复大量可用的复用机会。

由于每个令牌最多只会预填充一次，对公式（9求和后可得到累计证据令牌的上界：

$$\sum_{t=1}^{T} \sum_i p_i^{(t)} = \sum_i c_i^{(T)} = \sum_i \max_{t \le T} b_i^{(t)} \le \sum_i b_i^{\max}. \tag{10}$$

注意力细化。 多轮对话还提供了另一种独特机会：利用

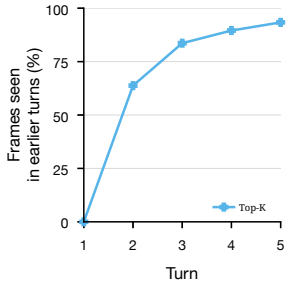

图 6. 后续轮次常会引用先前轮次中已预填充的帧。

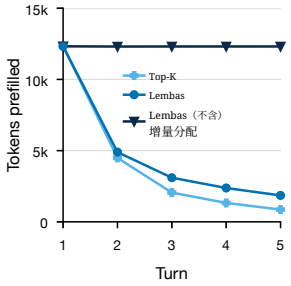

图 7. 增量分配使 Lembas 能够恢复大量可用的复用机会。

视觉语言模型对先前缓存证据的注意力来细化相关性。由于无论如何都必须预填充查询，Lembas 会在分配前预填充 q_t，从而无需额外推理成本即可获取查询词元对每个缓存项目的注意力。令 $a_i^{(t)}$ 表示项目 i 接收到的注意力，并按每个缓存词元进行归一化。由于未缓存项目没有可用的注意力，我们只细化先前见过的证据的相关性：

$$\hat{r}_i^{(t)} = \begin{cases} \alpha_t\, r_i(q_t)^{1-\omega} (a_i^{(t)})^{\omega}, & i \in \mathcal{C}_{t-1}, \\ r_i(q_t), & \text{otherwise}, \end{cases} \tag{11}$$

其中，$\mathcal{C}_{t-1}$ 是缓存项目的集合，$\omega \in [0,1]$ 控制注意力的贡献，而 α_t 则对融合分数进行归一化，以保留缓存项目的总检索相关性。因此，注意力会重新排序先前观测到的证据，但不会改变未见候选项的相关性分数。值得注意的是，细化过程只更新项目相关性，而校准后的令牌价值函数 $v_i(b_i)$保持不变。随后，令牌分配器使用$\hat{r}_i^{(t)}$ 在公式（2）中，仅对分配差异进行增量预填充。

4. 评估

我们在两种视觉语言模型骨干网络上评估 Lembas：Qwen3-VL-8B [2] 和 InternVL3.5-4B [47]，，并默认使用 Qwen3-VL-8B。作为一种即插即用的令牌分配框架，Lembas 可与现有的视频检索增强生成系统互补，而非取而代之。我们在两种检索流程中评估其有效性：(i) 基于嵌入的检索增强生成，该流程以 1 帧/秒的频率采样视频帧，并与时间对齐字幕一同建立索引，使用 SigLIP 2 [46]和 BGE [53]分别计算它们与查询的相关性；(ii) 基于图的检索增强生成，该流程基于 VideoRAG [38], 构建，通过图结构检索相关的视觉和文本证据。

基准测试。 我们在三个广泛使用的单轮、多项选择视频理解基准以及一个多轮开放式基准上进行评估：

方法	MLVU（不含字幕）		LVB（不含字幕）		Video-MME（含字幕）				
	万亿次浮点运算 ↓	总体 ↑	万亿次浮点运算 ↓	总体 ↑	万亿次浮点运算 ↓	总体 ↑	短 ↑	中等 ↑	长 ↑
Qwen3-VL-8B [2]									
均匀采样	215	73.4	229	65.1	224	72.9	**83.6**	73.0	62.0
RAG									
使用 Top-K	218	75.9	232	64.0	230	73.1	82.4	73.4	63.6
使用 Q-Frame [62]$_{\text{ICCV'25}}$	209	74.7	225	65.7	218	71.1	80.1	70.3	63.0
使用 QuoTA [33]$_{\text{AAAI'26}}$	781	77.2	676	**66.6**	647	74.4	83.2	74.4	65.4
使用 FOCUS [65]$_{\text{ICLR'26}}$	211	71.7	226	64.0	224	70.1	79.1	69.4	61.8
使用 AdaAlloc [1]$_{\text{NeurIPS'26}}$	203	73.2	222	65.1	225	71.0	82.8	69.4	60.7
使用 **Lembas（我们的方法）**	230	**78.1**	234	**66.6**	232	**77.3**	**83.6**	**76.8**	**71.7**
VideoRAG [38]$_{\text{KDD'26}}$	211	73.8	207	**61.9**	220	72.3	**81.0**	74.0	61.8
使用 **Lembas（我们的方法）**	208	**74.2**	200	61.2	217	**74.5**	80.5	**74.9**	**68.0**
InternVL3.5-4B [47]									
均匀采样	143	69.9	143	62.0	141	65.6	**79.8**	63.8	53.2
RAG									
使用 Top-K	145	72.5	146	60.4	145	68.6	78.0	69.1	58.7
使用 Q-Frame [62]$_{\text{ICCV'25}}$	146	71.1	146	63.1	142	66.0	76.7	64.6	56.9
使用 QuoTA [33]$_{\text{AAAI'26}}$	708	71.3	583	62.5	551	64.9	74.7	64.6	55.4
使用 FOCUS [65]$_{\text{ICLR'26}}$	142	69.5	143	58.8	140	66.4	75.9	65.9	57.4
使用 AdaAlloc [1]$_{\text{NeurIPS'26}}$	140	69.6	143	61.3	139	65.7	76.3	66.4	54.2
与 **Lembas（我们的方法）**	152	**73.0**	152	**63.2**	150	**72.1**	79.0	**70.8**	**66.4**

表 1. 准确率在三个视频理解基准测试上，以匹配的计算量进行比较，涵盖所有基线方法和两种骨干网络 es.

•**MLVU** [64]：涵盖来自 1,242 个视频的 2,174 道选择题，视频时长从 3 分钟到 2 小时不等，涉及从主题推理到动作计数的七项任务。

•**LongVideoBench（LVB）** [51]：包含 1,337 道选择题，涉及 752 个视频，时长从 8 秒到 1 小时，涵盖 17 项感知与关系任务。

•**Video-MME** [10]：包含 2,700 道选择题，涉及 900 个视频，分为短（<2 分钟）、中等（4–15 分钟）和长（30–60 分钟）子集。

• **MT-Video-Bench（MTVB）** [36]：包含 1,000 段开放式对话（5–8 轮），涉及 135 个视频，时长从 20 秒到 41 分钟。开放式回答依据给定的答案核对清单，由 LLM 评审器（Gemini3.8 Flash）评分。

基线方法。我们将 Lembas 与均匀采样以及近期的五种帧选择和令牌分配方法进行比较，所有方法均在相同的令牌预算下运行。

• **Top-K**：检索与查询的帧语义相似度得分最高的 K 个帧。

• **FOCUS** [65]：通过在时间片段上进行多臂老虎机探索来选择 K 个帧。

• **Q-Frame** [62]：根据查询相关性采样帧，并按照 2:1:1 的预算比例，在高、中、低分辨率层级之间分配词元。

• **QuoTA** [33]：将查询改写为相关实体，并使用视觉语言模型根据每帧包含这些实体的可能性为其评分。随后按比例根据评分分配词元，并将每帧的词元下采样至特定分辨率。

• **AdaAlloc** [1]：使用智能体规划如何在低分辨率和高分辨率帧层级之间分配预算。

对于多轮评估，当某种方法选择的证据与先前预填充的内容完全匹配时，我们启用 KV缓存复用：Top-K 和 FOCUS 复用之前出现过的帧，而 Q-Frame 和 AdaAlloc 仅在分辨率匹配时复用帧。QuoTA 不支持这种复用，因为它会将词元下采样为数量不固定的形式，从而无法实现复用。

指标。我们报告回答准确率和推理计算量（万亿次浮点运算），后者通过令牌预算进行控制。为确保单轮与多轮调用方法之间的公平比较，我们统计所有模型调用的预填充计算，包括辅助帧评分和规划调用。离线索引（包括帧编码和令牌排序）在每个视频上仅运行一次，并由所有查询共享，因此我们对所有方法均不计入这部分计算量。实验在 A100 GPU 上进行。其他实现细节、超参数和敏感性分析见附录。

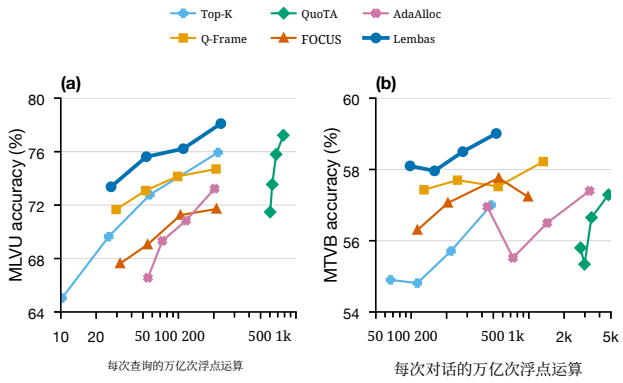

图 8. 在 (a) 单轮和 (b) 多轮视频理解任务中，Lembas 均实现了计算量与准确率之间的帕累托前沿。

方法	万亿次浮点运算 ↓	Video-MME（含字幕） ↑			
		总体	短	中等	Long
Lembas	232	**77.3**	**83.6**	**76.8**	**71.7**
无文本分配	226	73.7	81.9	74.9	64.4
无多样性项	226	76.6	82.4	76.4	70.8
无注水法	229	76.1	82.3	76.1	69.9

表2. 单轮问答中的消融实验。

方法	万亿次浮点运算 ↓	MTVB ↑
Lembas	275	58.5
无注意力细化	315	57.4
无增量分配	687	**59.1**

表3. 多轮问答的消融实验。

4.1. 主要结果

Lembas 在单轮视频理解中实现了帕累托前沿。 如表 1所示，Lembas 在三个基准和两种视觉语言模型骨干网络上均取得了最高的总体准确率。使用 Qwen3-VL-8B 时，在计算量相当的情况下，Lembas 在 MLVU、LVB 和 Video-MME 上分别较 Top-K 提高了 2.2、2.6 和 4.2 个百分点。在长视频上，这一提升尤为显著：在 Video-MME 的长视频子集上，Lembas 将准确率从 63.6% 提高到 71.7%，而万亿次浮点运算几乎相同。QuoTA 取得了具有竞争力的准确率，但由于采用基于视觉语言模型的帧评分，其计算量高出 2.8–34× 倍。这些趋势同样适用于 InternVL3.5-4B：Lembas 较 Top-K 提高了 0.5–3.5 个百分点。此外，在多个令牌预算下，Lembas 始终实现了有利的准确率—计算量权衡（图 8a）。

与 VideoRAG 集成后，Lembas 通过在更多帧之间分配令牌，在计算量相近的情况下，使 MLVU 和 Video-MME 上的准确率分别提高了 0.4 和 2.2 个百分点，而 LVB 上的准确率则下降了 0.7 个百分点。这些结果表明，Lembas 可与现有 RAG设计互为补充。

Lembas 在多轮视频理解任务中实现了帕累托前沿。 如图8b 所示，Lembas 的表现优于所有基线方法，准确率高出 0.3–1.5 个百分点，同时所需万亿次浮点运算减少了 1.7–17 倍。3× 与以往仅复用先前选定帧或分辨率完全匹配内容的方法不同，Lembas 仅对 KV缓存中尚不存在的词元进行增量预填充。这种更细粒度的复用大幅减少了多轮对话中的冗余计算，同时保持了答案准确率。

Lembas 带来的计算量开销可以忽略不计。使用 FAISS 进行语义相似性搜索、 [24], 迭代注水算法和令牌剪枝，耗时总计约为几十毫秒，远少于视觉语言模型的推理时间。

性能分析。我们对表格中的 Lembas 关键组件进行了消融。2 和 3。在 Video-MME 上，移除字幕分配会使总体准确率下降 3.6 个百分点，在长视频上的降幅尤其明显，达 7.3 个百分点。即使没有字幕分配，Lembas 的表现仍比预算相近的所有非智能体帧选择基线方法高出 0.6–3.6 个百分点。移除多样性项会使准确率下降 0.7 个百分点，证实了对冗余进行建模的益处；将注水法替换为更简单的比例分配方法（该方法无法提供最优性保证），则会使准确率下降 1.2 个百分点。

在多轮对话中，移除注意力细化会使准确率下降 1.1 个百分点，同时将预填充计算量从 275 万亿次浮点运算提高到 315 万亿次浮点运算。禁用增量分配会使计算量从 275 万亿次浮点运算增加到 687 万亿次浮点运算（2.5×），但准确率仅提高 0.6 个百分点。这些结果表明，注意力细化能够改善证据选择，而增量分配是实现多轮计算节省的主要因素。

5. 结论

我们提出了 Lembas，这是一种令牌高效视频检索增强生成的即插即用框架，可在帧和字幕之间联合分配共享令牌预算。其令牌分配器通过迭代注水法优化查询相关性、令牌价值递减和证据多样性，同时通过增量分配和注意力细化实现跨对话轮次的高效复用。在计算量相当的情况下，Lembas 的准确率比当前最先进基线提高了 0.1–7.2 个百分点。在多轮视频理解任务中，Lembas 的准确率提高了 0.3–1.5 个百分点，同时减少了 1.7–17.3× 万亿次浮点运算。

大语言模型使用声明

在准备这项工作时，我们获得了多个大语言模型的协助，具体来自 ChatGPT、Claude 和 Gemini 模型系列。这些模型用于协助论文写作、相关工作检索和编程，作者对相关内容进行了审阅和编辑。在核心思想和实验的设计与分析方面，大语言模型的使用程度较低。我们使用 Gemini 3.8 Flash 评估了多轮基准测试结果。我们未使用大语言模型创建或标注数据。作者对全部内容承担全部责任。

参考文献

[1] Joungbin An 和 Kristen Grauman。AdaAlloc：面向长视频问题回答的自适应视觉令牌分配。载于《神经信息处理系统进展》（NeurIPS），2026 年。1、3、7、16[2] Shuai Bai、Yuxuan Cai、Ruizhe Chen、Keqin Chen、Xionghui Chen、Zesen Cheng、Lianghao Deng、Wei Ding、Chang Gao、Chunjiang Ge 等。Qwen3-VL 技术报告。arXiv 预印本 arXiv:2511.21631，2025 年。1、6、7[3] Suresh Bolusani、Mathieu Besanc¸on、Ksenia Bestuzheva、Antonia Chmiela、Jo˜ao Dion´ısio、Tim Donkiewicz 等。SCIP 优化套件 9.0。技术报告，Optimization Online，2024 年。16[4] Stephen Boyd 和 Lieven Vandenberghe。《凸优化》。剑桥大学出版社，2004 年。5、13[5] Shyamal Buch、Arsha Nagrani、Anurag Arnab 和 Cordelia Schmid。用于高效视频推理的灵活帧选择。载于 CVPR，2025 年。1[6] Liang Chen、Haozhe Zhao、Tianyu Liu、Shuai Bai、Junyang Lin、Chang Zhou 和 Baobao Chang。第 2 层之后，一张图像只需 1/2 个词元：面向大型视觉语言模型的即插即用推理加速。载于 ECCV，2024 1, 3。[7] Mung Chiang、Steven H. Low、A. Robert Calderbank 和 John C. Doyle。将分层视为优化分解：网络架构的数学理论。《IEEE 会议录》，95(1)：255–312，2007 年。3[8] Steven Diamond 和 Stephen Boyd。CVXPY：一种用于凸优化的 Python 嵌入式建模语言。《机器学习研究期刊》，17 83 1 5, 2016 16 ( )：–。[9] Marguerite Frank 和 Philip Wolfe。一种二次规划算法。《海军研究后勤季刊》，3(1–2)：95–110，1956 年。5[10] Chaoyou Fu、Yuhan Dai、Yongdong Luo、Lei Li、Shuhuai Ren、Renrui Zhang、Zihan Wang、Chenyu Zhou、Yunhang Shen、Mengdan Zhang 等。Video-MME：视频分析中多模态大语言模型的首个全面评估基准测试。载于 CVPR，2025 年。3、7[11] Tianyu Fu、Tengxuan Liu、Qinghao Han、Guohao Dai、Shengen Yan、Huazhong Yang、Xuefei Ning 和 Yu Wang。FrameFusion：结合相似性与重要性，在大型视觉语言模型上实现视频词元缩减。载于 ICCV，2025 年。1[12] Michael R. Garey 和 David S. Johnson。《计算机与难解性：NP 完全性理论指南》。W. H. Freeman and Company，旧金山，1979 年。12[13] Ali Ghodsi、Matei Zaharia、Benjamin Hindman、Andy Kon- winski、Scott Shenker 和 Ion Stoica。主导资源公平：多种资源类型的公平分配。载于 USENIX 网络系统设计与实现研讨会（NSDI），2011。3[14] Andrea J. Goldsmith 和 Pravin P. Varaiya。具有信道侧信息的衰落信道容量。《IEEE 信息论汇刊》，43(6):1986–1992，1997。3[15] Paul J. Goulart 和 Yuwen Chen。Clarabel：用于具有二次目标函数的锥规划的内点求解器。arXiv 预印本 arXiv:2405.12762，2024。16[16] Robert Grandl、Ganesh Ananthanarayanan、Srikanth Kan- dula、Sriram Rao 和 Aditya Akella。面向集群调度器的多资源装箱。载于 ACM SIGCOMM 会议，第455–466页，2014。3[17] Gurobi Optimization, LLC。Gurobi 优化器参考手册，2026。4，16[18] Benjamin Hindman、Andy Konwinski、Matei Zaharia、Ali Ghodsi、Anthony D. Joseph、Randy Katz、Scott Shenker 和 Ion Stoica。Mesos：数据中心细粒度资源共享平台。载于 USENIX 网络系统设计与实现研讨会（NSDI），2011。3[19] Kevin Hsieh、Ganesh Ananthanarayanan、Peter Bodik、Shiv- aram Venkataraman、Paramvir Bahl、Matthai Philipose、Phillip B. Gibbons 和 Onur Mutlu。FOCUS：以低延迟、低成本查询大型视频数据集。载于 USENIX 操作系统设计与实现研讨会（OSDI），第269–286页，2018。5[20] IBM ILOG CPLEX。V12.1：CPLEX 用户手册。国际商业机器公司，2009。4[21] Michael Isard、Vijayan Prabhakaran、Jon Currey、Udi Wieder、Kunal Talwar 和 Andrew Goldberg。Quincy：分布式计算集群的公平调度。载于 ACM SIGOPS 操作系统原理研讨会（SOSP），第261–276页，2009。3[22] Martin Jaggi。重新审视 Frank-Wolfe：无投影稀疏凸优化。载于国际机器学习会议（ICML），第427–435页，2013。5，14[23] Soyeong Jeong、Kangsan Kim、Jinheon Baek 和 Sung Ju Hwang。VideoRAG：对视频语料库进行检索增强生成。载于 ACL论文集，2025。2[24] Jeff Johnson、Matthijs Douze 和 Herv´e J´egou。使用 GPU 进行十亿规模的相似性搜索。《IEEE 大数据汇刊》，7(3):535–547，2021。8[25] Richard M. Karp。组合问题之间的可归约性。载于《计算机计算复杂性》，第85–103页。Springer US，1972。12[26] Frank P. Kelly、A. K. Maulloo 和 D. K. H. Tan。通信网络的速率控制：影子价格、比例公平与稳定性。《运筹学会志》，49(3):237–252，1998。3[27] Hosu Lee、Junho Kim、Hyunjun Kim 和 Yong Man Ro。Re-FoCUS：强化学习引导的帧优化，用于上下文理解。载于 CVPR Findings，2026。1[28]Patrick Lewis、Ethan Perez、Aleksandra Piktus、Fabio Petroni、Vladimir Karpukhin、Naman Goyal、Heinrich

K¨uttler、Mike Lewis、Wen-tau Yih、Tim Rockt¨aschel、Sebastian Riedel 和 Douwe Kiela。面向知识密集型自然语言处理任务的检索增强生成。载于 NeurIPS，2020。2

[29] Jialuo Li、Bin Li、Jiahao Li 和 Yan Lu。先划分，再对齐：针对查询类型调整帧选择，以实现长篇视频理解。载于 CVPR，2026。2[30] Shuming Liu、Chen Zhao、Tianqi Xu 和 Bernard Ghanem。BOLT：无需训练即可增强大型视觉语言模型，实现长篇视频理解。载于 CVPR，2025。2[31] Steven H. Low 和 David E. Lapsley。优化流量控制—I：基本算法与收敛性。《IEEE/ACM网络学报》，7(6):861–874，1999。3[32] Yongdong Luo、Xiawu Zheng、Guilin Li、Shukang Yin、Haojia Lin、Chaoyou Fu、Jinfa Huang、Jiayi Ji、Fei Chao、Jiebo Luo 和 Rongrong Ji。Video-RAG：视觉对齐的检索增强长视频理解。载于 NeurIPS，2025。1，2[33] Yongdong Luo、Wang Chen、Weizhong Huang、Shukang Yin、Haojia Lin、Jinfa Huang、Chaoyou Fu、Jiayi Ji、Xiawu Zheng 和 Jiebo Luo。QuoTA：通过思维链查询解耦进行面向查询的令牌分配，以实现长视频理解。载于 AAAI 人工智能会议论文集，第24160–24168页，2026。1，3，5，7，14，16[34] Ziyu Ma、Chenhui Gou、Hengcan Shi、Bin Sun、Shutao Li、Hamid Rezatofighi 和 Jianfei Cai。DrVideo：基于文档检索的长视频理解。载于 CVPR，2025 2。

[35] Daniel P. Palomar 和 Mung Chiang。网络效用最大化分解方法教程。《IEEE通信领域精选期刊》，24(8):1439–1451，2006。3[36] Yaning Pan、Qianqian Xie、Guohui Zhang、Zekun Moore Wang、Yongqian Wen、Yuanxing Zhang、Haoxuan Hu、Zhiyu Pan、Yibing Huang、Zhidong Gan、Yonghong Lin、An Ping、Shihao Li、Yanghai Wang、Tianhao Peng 和 Jiaheng Liu。MT-Video-Bench：用于评估多轮对话中多模态大语言模型的整体视频理解基准。载于 ACL 论文集，2026。7[37] Jun Peng、Baiyang Song、Jie Li、Hui Li、Yiyi Zhou、Rongrong Ji 和 Yonghong Tian。QCA：面向查询与内容的长视频理解关键帧选择。载于 ECCV，2026。1，2[38] Xubin Ren、Lingrui Xu、Long Xia、Shuaiqiang Wang、Dawei Yin 和 Chao Huang。VideoRAG：利用超长上下文视频进行检索增强生成。载于 ACM SIGKDD知识发现与数据挖掘会议（KDD），第2390 2401, 2026 1, 2, 6, 7 –页。

[39] Gesualdo Scutari、Daniel P. Palomar 和 Sergio Barbarossa。高斯频率选择性干扰信道的异步迭代注水。《IEEE 信息论汇刊》，54(7):2868–2878，2008。3[40] Kele Shao、Keda Tao、Can Qin、Haoxuan You、Yang Sui 和 Huan Wang。HoliTom：面向快速视频大语言模型的整体令牌合并。载于 NeurIPS，2025。1[41] Xiaoqian Shen、Wenxuan Zhang、Jun Chen 和 Mohamed Elhoseiny。Vgent：面向长视频理解的图结构检索-推理增强生成。载于 NeurIPS，2025。1，2

[42] Hongchang Shi、Jinpeng Hu、Ao Wang、Wenzheng Zhou、Hui Ma、Feng Li 和 Zenglin Shi。《MarKey：边际效用引导的长视频理解贪心关键帧选择》。arXiv 预印本 arXiv:2609.15408，2026。1，2[43] Hui Sun、Shiyin Lu、Huanyu Wang、Qing-Guo Chen、Zhao Xu、Weihua Luo、Kaifu Zhang 和 Ming Li。《MDP3：一种用于视频大语言模型列表式帧选择的免训练方法》。发表于 ICCV，2025。[44] Xi Tang、Jihao Qiu、Lingxi Xie、Yunjie Tian、Jianbin Jiao 和 Qixiang Ye。《长视频理解的自适应关键帧采样》。发表于 CVPR，2025。1，3[45] Keda Tao、Can Qin、Haoxuan You、Yang Sui 和 Huan Wang。《DyCoke：用于快速视频大语言模型的令牌动态压缩》。发表于 CVPR，第 18992–19001 页，2025。1 – 。[46] Michael Tschannen、Alexey Gritsenko、Xiao Wang、Muhammad Ferjad Naeem、Ibrahim Alabdulmohsin、Nikhil Parthasarathy 等。《SigLIP 2：具备改进的语义理解、定位和密集特征的多语言视觉语言编码器》。arXiv 预印本 arXiv:2502 14786，2025。6，15。[47] Weiyun Wang 等。《InternVL3.5：提升开源多模态模型的通用性、推理能力与效率》。arXiv 预印本 arXiv:2508.18265，2025。1，6，7[48] Xiaohan Wang、Yuhui Zhang、Orr Zohar 和 Serena Yeung-Levy。《VideoAgent：以大语言模型为智能体的长篇视频理解》。发表于 ECCV，2024。1[49] Yilin Wang、Xiangxi Zheng、Dongxing Mao、Linjie Li、Zhengyuan Yang、Ping Yu、Rui Yan、Yuan Yao 和 Alex Jinping Wang。《利用基于注意力的多模态大语言模型选择器在测试时对长视频进行高效帧选择》。发表于 EMNLP 论文集，2026。1，3[50] Ziyang Wang、Shoubin Yu、Elias Stengel-Eskin、Jaehong Yoon、Feng Cheng、Gedas Bertasius 和 Mohit Bansal。《VideoTree：用于长视频大语言模型推理的自适应树状视频表示》。发表于 CVPR，2025 2。

[51] Haoning Wu、Dongxu Li、Bei Chen 和 Junnan Li。《LongVideoBench：长上下文交错视频-语言理解基准测试》。发表于神经信息处理系统进展（NeurIPS），数据集与基准赛道，2024。3，7[52] Guangxuan Xiao、Yuandong Tian、Beidi Chen、Song Han 和 Mike Lewis。《带注意力汇的高效流式语言模型》。发表于国际学习表征会议（ICLR），2024。3[53] Shitao Xiao、Zheng Liu、Peitian Zhang、Niklas Muennighoff、Defu Lian 和 Jian-Yun Nie。《C-Pack：用于通用中文嵌入的打包资源》。发表于 SIGIR，2024。6，15[54] Junlin Xie、Quanlong Zheng、Ruifei Zhang、Kuo Wang、Yan-hao Zhang、Jinguo Luo、Haonan Lu、Xiang Wan 和 Guan-bin Li。《StreamRAG：通过检索增强提升实时视频理解》。发表于 CVPR，2026。2[55] Yuxuan Yan、Shiqi Jiang、Ting Cao、Yifan Yang、Qianqian Yang、Yuanchao Shu、Yuqing Yang 和 Lili Qiu。《AVA：利用视觉语言模型实现智能体视频分析》。发表于 NSDI，2026。1，2[56] Senqiao Yang、Yukang Chen、Zhuotao Tian、Chengyao Wang、Jingyao Li、Bei Yu 和 Jiaya Jia。《VisionZip：更长

在视觉语言模型中，更长并非必要。发表于 CVPR，2025。1，3[57] Jinhui Ye、Zihan Wang、Haosen Sun、Keshigeyan Chandrasegaran、Zane Durante、Cristobal Eyzaguirre、Yonatan Bisk、Juan Carlos Niebles、Ehsan Adeli、李飞飞、Jiajun Wu 和 Manling Li。《重新思考长篇视频理解的时序搜索》。发表于 CVPR，2025。1[58] Shoubin Yu、Jaemin Cho、Prateek Yadav 和 Mohit Bansal。《用于视频定位和问答的自链式图像-语言模型》。发表于 NeurIPS，2023。1[59] Sicheng Yu、Chengkai Jin、Huanyu Wang、Zhenghao Chen、Sheng Jin、Zhongrong Zuo、Xiaolei Xu、Zhenbang Sun、Bingni Zhang、Jiawei Wu、Hao Zhang 和 Qianru Sun。《Frame-Voyager：学习为视频大语言模型查询帧》。发表于国际学习表征会议（ICLR），2025。1，3[60] Wei Yu、Wonjong Rhee、Stephen Boyd 和 John M. Cioffi。《高斯向量多址信道的迭代注水》。IEEE信息论汇刊，50(1)：145–152，2004。3，5[61] Peiyuan Zhang、Kaichen Zhang、Bo Li、Guangtao Zeng、Jingkang Yang、Yuanhan Zhang、Ziyue Wang、Haoran Tan、Chunyuan Li 和 Ziwei Liu。《从语言到视觉的长上下文迁移》。机器学习研究汇刊，2025。3[62] Shaojie Zhang、Jiahui Yang、Jianqin Yin、Zhenbo Luo 和 Jian Luan。《Q-Frame：视频大语言模型的查询感知帧选择与多分辨率适配》。发表于 IEEE/CVF 国际计算机视觉会议（ICCV），2025。1，3，5，7，14，15，16[63] Zhenyu Zhang、Ying Sheng、Tianyi Zhou、Tianlong Chen、Lianmin Zheng、Ruisi Cai、Zhao Song、Yuandong Tian、Christopher Re、Clark Barrett、Zhangyang Wang 和 Beidi Chen。《H2O：用于大语言模型高效生成式推理的重击者预言机》。发表于 NeurIPS，2023。3[64] Junjie Zhou、Yan Shu、Bo Zhao、Boya Wu、Zhengyang Liang、Shitao Xiao、Minghao Qin、Xi Yang、Yongping Xiong、Bo Zhang、Tiejun Huang 和 Zheng Liu。《MLVU：多任务长视频理解基准测试》。发表于 IEEE/CVF 计算机视觉与模式识别会议论文集（CVPR），2025。3，7[65] Zirui Zhu、Hailun Xu、Yang Luo、Yong Liu、Kanchan Sarkar、Zhenheng Yang 和 Yang You。《 FOCUS：长视频理解的高效关键帧选择》。发表于国际学习表征会议（ICLR），2026。1，3，7，15[66] Yuanhao Zou、Shengji Jin、Andong Deng、Youpeng Zhao、Jun Wang 和 Chen Chen。《A.I.R.：为视频问答实现自适应、迭代式和基于推理的帧选择》。发表于国际学习表征会议（ICLR），2026。1

附录

本附录给出了正文（§A）中省略的推导与证明、实现细节（§B）以及补充结果（§C）。

A. 推导与证明

为完整起见，我们首先重述正文中的分配问题。设 $\mathcal{F}$ 和 $\mathcal{S}$ 分别为候选帧和字幕，$b_i \in [0, b_i^{\max}]$ 为项目 $i \in \mathcal{F} \cup \mathcal{S}$ 的令牌数量，$v_i(b_i) = b_i/(b_i+\tau_i)$为其令牌价值，其中$\tau_i$ 对于帧为 τ_{frame}，对于字幕为 τ_{text}。Lembas 求解

$$\begin{aligned} \max_{\mathbf{b}} \quad & \sum_{i\in\mathcal{F}} r_i(q,\mathbf{b})\, v_i(b_i) + \beta \sum_{i\in\mathcal{S}} r_i(q,\mathbf{b})\, v_i(b_i) \\ \text{s.t.} \quad & \sum_{i\in\mathcal{F}\cup\mathcal{S}} b_i \le B, \end{aligned} \tag{12}$$

其中，依赖于分配的相关性 $r_i(q,\ \mathbf{b})$ 会根据项目 i 所属模态 $\mathcal{M}_i \in \{\mathcal{F}, \mathcal{S}\}$ 中相似项目的令牌价值，对其检索相关性 $r_i(q)$ 进行折减：

$$r_i(q,\mathbf{b}) = r_i(q) - \eta \sum_{j\in\mathcal{M}_i} s_{ij}\, v_j(b_j). \tag{13}$$

此处，s_{ij} 是项目 i 和 j嵌入之间的余弦相似度，$\eta \ge 0$ 是冗余惩罚。

A.1. 离散问题的 NP-困难性

视觉语言模型消耗完整令牌，因此在实践中，每个令牌数量 b_i 必须是 $\{0, 1, \ldots, b_i^{\max}\}$中的整数。我们证明，式（12）的这一离散版本是 NP-困难的。其判定版本要求判断预算 B 内是否存在某个整数分配，使目标值至少达到 θ。

命题 2。对式（12）进行最大化，在令牌数量为整数时是 *NP*-困难的，即使每个项目最多只能分配一个词元。

我们从 NP-完全的独立集问题进行归约 [12, 25]：给定一个图 $G = (V, E)$，其中有 n个顶点和一个整数 k，判断 G 是否包含 k 个顶点，且其中任意两个顶点都不相邻。其思想是，冗余惩罚充当对 G中的边的惩罚。每个顶点对应一个帧；两个相邻帧如果都获得一个词元，就会降低目标值，因此，当独立集存在时，最优的 k 帧构成一个独立集。

证明。构造。给定 G 和 k，我们构造一个实例：每个顶点对应一帧，且不含字幕：

$$\begin{aligned} & b_i^{\max} = 1, \quad \tau_i = 1, \quad r_i(q) = \tfrac{1}{n}, \\ & \eta = \tfrac{1}{n}, \quad B = k, \quad \mathbf{S} = \mathbf{I} + \tfrac{1}{n}\mathbf{A}, \end{aligned} \tag{14}$$

其中$\mathbf{A}$是 G的邻接矩阵。因此，两个相邻帧的相似度为 $1/n$，而其他所有不同帧对的相似度均为 0。该构造所需时间为多项式时间。

有效性。目标函数仅通过嵌入之间的余弦相似度来决定，因此只需保证$\mathbf{S}$ 是某些单位向量的余弦相似度矩阵。$\mathbf{A}$ 的每个特征值的绝对值都不超过 G的最大度数，而该最大度数至多为 $n-1$，因此$\mathbf{S}$ 的每个特征值都至少为 $1/n > 0$。因此，$\mathbf{S}$是正定且对角线为单位对角线的矩阵，也就是说，它是 n 个单位向量的 Gram 矩阵。

目标函数。每帧要么不分配令牌，其令牌价值为0，要么分配一个令牌，其令牌价值为 $v_i(1) = 1/2$。令 X 表示获分配令牌的帧集合，并令 $e(X)$表示 G 中两个端点均位于 X中的边数。预算要求 $|X| \le k$，而式（12）中的目标函数变为

$$\begin{aligned} J(X) &= \frac{|X|}{2n} - \frac{1}{4n} \sum_{i,j\in X} s_{ij} \\ &= \frac{|X|}{2n} - \frac{|X|}{4n} - \frac{e(X)}{2n^2} = \frac{|X|}{4n} - \frac{e(X)}{2n^2}. \end{aligned} \tag{15}$$

正确性。如果 G包含一个大小为k的独立集，那么给其中每一帧分配一个令牌是可行的，并且可达到 $J = k/(4n)$。反过来，每一种可行分配都满足 $J(X) \le |X|/(4n) \le k/(4n)$，且仅当 $|X| = k$ 且 $e(X) = 0$时等号成立，也就是说，仅当 X 是一个大小为 k的独立集时等号成立。因此，G包含一个大小为k 的独立集，当且仅当离散问题达到 $\theta = k/(4n)$。由此，若存在一个求解离散问题的多项式时间算法，就能在多项式时间内判定独立集问题。

该构造依赖于冗余惩罚项，它将各个项目耦合起来：若没有该惩罚项（$\eta = 0$），给任意 k 帧分配一个令牌都是最优的。在连续松弛中，同样的耦合作用仍会得到一个凸规划，因为 $\mathbf{S}$ 是半正定的（§A.3）。因此，Lembas 求解该松弛问题，并将其解向下取整为完整令牌。

A.2. 水填充推导

我们将正文中的注水分配推导为以下子问题的唯一解，其中每个项目的相关性固定为 $\bar{r}_i > 0$:

$$\begin{aligned} \max_{\mathbf{b}} \quad & \sum_i \bar{r}_i \frac{b_i}{b_i + \tau_i} \\ \text{s.t.} \quad & \sum_i b_i \le B, \quad 0 \le b_i \le b_i^{\max}. \end{aligned} \tag{16}$$

我们通过该子问题的拉格朗日对偶来求解。其思路是将预算替换为每个词元的价格，让每个项目按该价格获取词元，然后不断调整价格，直到预算用尽。我们为预算约束引入乘子 $\lambda \ge 0$，并将边界 $0 \le b_i \le$

$b_i^{\max}$作为 $\mathbf{b}$ 的定义域，由此得到拉格朗日函数

$$\mathcal{L}(\mathbf{b},\lambda)=\sum_i \bar{r}_i\,\frac{b_i}{b_i+\tau_i}-\lambda\Big(\sum_i b_i-B\Big). \tag{17}$$

该子问题等价于 最大化$_{\mathbf{b}}$ 最小化$_{\lambda\ge 0}$ $\mathcal{L}(\mathbf{b}\ \lambda)$，其对偶问题为 最小化$_{\lambda\ge 0}$ 最大化$_{\mathbf{b}}$ $\mathcal{L}(\mathbf{b}\ \lambda)$。由于目标函数是凹函数、约束是线性的，且 $\mathbf{b}=\mathbf{0}$ 严格可行 [4]，因此强对偶性成立。因此，这两个问题取得相同的值，且子问题的解在对偶最优解 $\mathcal{L}(\cdot,\ \lambda^\star)$ 处最大化 $\lambda^\star$。

内部最大化。 对于固定的 $\lambda>0$，拉格朗日函数按项目分离，并且每个项目求解

$$\max_{0\le b_i\le b_i^{\max}}\ \bar{r}_i\,\frac{b_i}{b_i+\tau_i}-\lambda\, b_i. \tag{18}$$

该目标函数是严格凹函数。令其导数为零，可得

$$\frac{\bar{r}_i\,\tau_i}{(b_i+\tau_i)^2}=\lambda, \tag{19}$$

将解截断到定义域后，得到唯一最大化解

$$b_i(\lambda)=\Big[\sqrt{\bar{r}_i\tau_i/\lambda}-\tau_i\Big]_0^{b_i^{\max}}, \tag{20}$$

这就是正文中的注水分配。直观地说，每个项目会持续获取词元，直到其边际增益降至价格 λ。因此，当某个项目在零个词元时的边际增益 $\bar{r}_i/\tau_i$ 不超过 λ 时，它不会获得任何词元；而当其在获得全部令牌数量 $b_i^{\max}$时的边际增益仍至少为 λ 时，它会获得全部词元。在迭代注水中，某个项目的折扣相关性可能为非正。此类项目无法增加目标函数，因此不会获得词元。

外部最小化。 将公式（20）代入公式（17）可得对偶函数 g（λ） $=\mathcal{L}$（$\mathbf{b}$（λ） λ），它是单变量 λ 的凸函数，其导数为

$$g'(\lambda)=B-\sum_i b_i(\lambda). \tag{21}$$

如果 $\sum_i b_i^{\max}\le B$，则预算不受约束，每个项目都会获得其完整的令牌数量。否则，对偶最优解$\lambda^\star$ 满足 $g'(\lambda^\star)$ $=0$，即分配恰好用完预算。由于存在截断，该方程通常没有闭式解。然而，每个 $b_i(\lambda)$关于 λ连续且非递增，并且当 $\lambda\ge\max_i \bar{r}_i/\tau_i$时它们都会变为零。因此，在 $\lambda\in(0,\max_i \bar{r}_i/\tau_i]$上进行二分法即可找到 $\lambda^\star$。由于内层最大化器是唯一的，由此得到的分配 $\mathbf{b}(\lambda^\star)$ 是方程（16）的唯一解。

A.3. 迭代注水的收敛性

现在我们证明正文中的命题1，该命题界定了迭代注水与方程（12）全局最大值之间的差距。我们首先证明连续松弛是一个凸规划，然后证明该界限。

连续松弛的凸性。 连续松弛允许每个令牌数量 b_i 在 $[0,b_i^{\max}]$中取任意实数值。经过变量替换后，该松弛是在凸集上最大化凹函数。

我们将变量从令牌数量改为令牌价值 $z_i=v_i(b_i)$，并用$\mathbf{b}$和$\mathbf{z}$ 表示所有项目上的向量。由于 v_i 严格递增，这一变量替换是可逆的，即

$$b_i=v_i^{-1}(z_i)=\frac{\tau_i z_i}{1-z_i}, \tag{22}$$

其中 $0\le z_i\le z_i^{\max}=v_i(b_i^{\max})\ <1$。在这些变量下，预算 B 将令牌价值限制在可行集

$$\mathcal{Z}=\Big\{\mathbf{z}:\ 0\le z_i\le z_i^{\max},\ \sum_i\frac{\tau_i z_i}{1-z_i}\le B\Big\}. \tag{23}$$

每个函数 $\tau_i z_i/(1-z_i)$ 在 $[0,1)$ 上是凸函数，因为其二阶导数 $2\tau_i/(1-z_i)^3$ 为正。预算约束因此将凸函数限制在 B以下，所以 $\mathcal{Z}$ 是凸的；由于它是闭且有界的，因此也是紧的。

反过来，目标函数关于令牌价值是凹函数。将式（13）代入式（12）可得

$$J(\mathbf{z})=\bar{\mathbf{r}}^\top\mathbf{z}-\eta\,\mathbf{z}^\top\mathbf{S}\mathbf{z}, \tag{24}$$

其中，帧对应 $\bar{r}_i=r_i(q)$，字幕对应 $\bar{r}_i=\beta\, r_i(q)$；相似度矩阵 $\mathbf{S}$ 为块对角矩阵，其各块分别为帧相似度矩阵和经 β 缩放的字幕相似度矩阵。每个块都包含同一模态嵌入之间的余弦相似度，包括对角线元素 $s_{ii}=1$。因此，帧块是归一化嵌入的 Gram 矩阵，字幕块则是此类矩阵的正倍数，所以 $\mathbf{S}$ 是半正定的。由于 $\eta\ge 0$， J 是凹函数。在 $\mathcal{Z}$ 上最大化 J 因而是一个凸规划，其中每个局部最大值都是全局最大值 [4]。

命题1（重述）。 设 $J^\star$ 为式（12）在连续令牌分配上的全局最大值，并设 $J^{(\ell)}$ 为主文中算法 *1* 经过 ℓ 次迭代后的目标值。那么

$$J^\star-J^{(\ell)}\le\frac{4\,\eta\,\sigma_{\max}D^2}{\ell+2}=O\Big(\frac{1}{\ell}\Big), \tag{25}$$

其中， $\sigma_{\max}$ 是帧与 β缩放后的字幕相似度矩阵的最大特征值， D 是在预算 B下可行令牌价值集合的直径。

证明遵循 Frank–Wolfe 方法的标准分析，[22]，并将其应用于令牌价值$\mathbf{z}$。证明使用梯度 $\nabla J(\mathbf{z}) = -\mathbf{r} - 2\eta \mathbf{S}\mathbf{z}$，即正文公式 (6) 中的折扣相关性 $-\mathbf{r}$，以及常数 $C = 2\eta\sigma_{\max}$，该常数用于界定 J的曲率。直径 D是 $\mathcal{Z}$中两点之间的最大距离。证明依赖两个引理。第一个引理界定当前令牌价值与最优值之间的差距，第二个引理界定一次迭代对目标函数的改善幅度。

引理1（距最优值）。对于每个 $\mathbf{z} \in \mathcal{Z}$，

$$J^\star - J(\mathbf{z}) \le \max_{\mathbf{y}\in\mathcal{Z}} \nabla J(\mathbf{z})^\top(\mathbf{y}-\mathbf{z}). \tag{26}$$

证明。令$\mathbf{z}^{\text{opt}}$ 为在 $\mathcal{Z}$上使 J 达到最大值的点。由于 J 是凹函数，其在$\mathbf{z}$ 处的线性近似位于函数图像之上，因此 $J^\star = J(\mathbf{z}^{\text{opt}}) \le J(\mathbf{z}) + \nabla J(\mathbf{z})^\top(\mathbf{z}^{\text{opt}} - \mathbf{z})$。右侧至多为所有 $\mathbf{y} \in \mathcal{Z}$上的最大值。

直观地说，线性近似所预测的最佳改进永远不会小于剩余差距。一次注水迭代恰好解决了这一最大化问题：以令牌价值表示，带有相关性的公式（16）的目标函数 $\nabla J(\mathbf{z})^\top\mathbf{y}$，其约束定义了 $\mathcal{Z}$。

引理2（单步进展）。对于每个 $\mathbf{z}$，$\mathbf{s} \in \mathcal{Z}$ 和$\gamma \in [0,1]$，

$$J\big(\mathbf{z}+\gamma(\mathbf{s}-\mathbf{z})\big) \ge J(\mathbf{z}) + \gamma\,\nabla J(\mathbf{z})^\top(\mathbf{s}-\mathbf{z}) - \frac{C}{2}\gamma^2 D^2. \tag{27}$$

证明。令$\mathbf{d} = \gamma(\mathbf{s}-\mathbf{z})$。由于 J是二次函数，其二阶展开是精确的：$J(\mathbf{z}+\mathbf{d}) = J(\mathbf{z}) + \nabla J(\mathbf{z})^\top\mathbf{d} - \eta\,\mathbf{d}^\top\mathbf{S}\mathbf{d}$。最后一项至少为 $-\eta\sigma_{\max}\|\mathbf{d}\|^2 = -\frac{C}{2}\|\mathbf{d}\|^2$，且 $\|\mathbf{d}\| \le \gamma D$。

也就是说，沿着 $\mathbf{s}$ 方向迈出大小为 γ 的一步，至少可以获得完整步长所预测改进量的 γ，减去一个随 γ^2缩小的曲率项。

命题1的证明。我们首先推导剩余差距的递推关系，然后通过归纳法求解。设 $\mathbf{z}^{(\ell-1)}$为第 ℓ 次迭代之前的令牌价值，其中 $\mathbf{z}^{(0)} = \mathbf{0}$，并设 $h_\ell = J^\star - J(\mathbf{z}^{(\ell)})$ 为剩余差距。第 ℓ 次迭代采用折扣相关性$\nabla J(\mathbf{z}^{(\ell-1)})$执行注水法，返回式（26）右侧的最大化解 $\mathbf{s}$（主文中为 $\mathbf{z}^\star$），并移动至 $\mathbf{z}^{(\ell)} = \mathbf{z}^{(\ell-1)} + \gamma_\ell(\mathbf{s} - \mathbf{z}^{(\ell-1)})$，其中 $\gamma_\ell = 2/(\ell+1)$。由于 $\mathcal{Z}$ 是凸的，该点属于 $\mathcal{Z}$。

应用引理2 并结合$\mathbf{z} = \mathbf{z}^{(\ell-1)}$ 和 $\gamma = \gamma_\ell$ 可得

$$\begin{aligned} J(\mathbf{z}^{(\ell)}) &\ge J(\mathbf{z}^{(\ell-1)}) + \gamma_\ell\,\nabla J(\mathbf{z}^{(\ell-1)})^\top(\mathbf{s}-\mathbf{z}^{(\ell-1)}) \\ &\quad - \frac{C}{2}\gamma_\ell^2 D^2 \\ &\ge J(\mathbf{z}^{(\ell-1)}) + \gamma_\ell\, h_{\ell-1} - \frac{C}{2}\gamma_\ell^2 D^2, \end{aligned}$$

其中第二个不等式用到了引理1。将两边同时减去 $J^\star$ 可得

$$h_\ell \le (1-\gamma_\ell)\,h_{\ell-1} + \frac{C}{2}\gamma_\ell^2 D^2. \tag{28}$$

因此，每次迭代都会消除剩余差距的 γ_ℓ，并增加一个阶为 γ_ℓ^2 的曲率项。

现在通过归纳法证明 $h_\ell \le 2CD^2/(\ell+2)$。对于$\ell = 1$，步长为 $\gamma_1 = 1$，根据式（28）可得 $h_1 \le CD^2/2 \le 2CD^2/3$。对于 $\ell \ge 2$，假设命题对 $\ell-1$成立，即 $h_{\ell-1} \le 2CD^2/(\ell+1)$。那么

$$\begin{aligned} h_\ell &\overset{\text{(a)}}{\le} \frac{\ell-1}{\ell+1}\,h_{\ell-1} + \frac{2CD^2}{(\ell+1)^2} \\ &\overset{\text{(b)}}{\le} \frac{\ell-1}{\ell+1}\cdot\frac{2CD^2}{\ell+1} + \frac{2CD^2}{(\ell+1)^2} \\ &= \frac{2CD^2\,\ell}{(\ell+1)^2} \overset{\text{(c)}}{\le} \frac{2CD^2}{\ell+2}, \end{aligned}$$

其中，(a) 将 $\gamma_\ell = 2/(\ell+1)$ 代入式（28），(b) 应用归纳假设，(c) 使用 $\ell(\ell+2) \le (\ell+1)^2$。将 $C = 2\eta\sigma_{\max}$ 代入后得到式（25），证明完毕。

命题1有两个推论。首先，没有冗余惩罚项（$\eta = 0$）时，上界为零，因此第一次迭代（即普通注水法）已经是最优的。其次，由于每个令牌价值都属于 $[0,1)$，直径满足 $D^2 \le |\mathcal{F}| + |\mathcal{S}|$，因此上界的分子至多为 $4\,\eta\,\sigma_{\max}(|\mathcal{F}| + |\mathcal{S}|)$。该上界适用于连续分配，Lembas 随后会将其向下取整为完整令牌。

A.4. 先前分配规则的次优性

正文指出，无论是比例分配 [33]还是分层分配 [62]，都无法最大化公式（16）中的固定相关性目标。本节将证明这两个结论。

引理3（比例分配）。假设至少有三个项目具有两两不同的相关性 $\bar{r}_i > 0$，并且共享同一个参数 τ；令 $b_i = c\,\bar{r}_i$ 满足 $c > 0$ 以下条件$0 < b_i < b_i^{\max}$。对于这些项目，该分配无法最大化公式（16）。

证明。根据第§A.2节，最优分配要求所有满足 $0 < b_i < b_i^{\max}$的项目具有相同的边际增益，即$\bar{r}_i\tau/(c\,\bar{r}_i + \tau)^2 = \lambda$对某个 $\lambda > 0$成立。重新整理该方程可知，每个这样的相关性都是二次方程 $\lambda c^2 x^2 + (2\lambda c\tau - \tau)x + \lambda\tau^2 = 0$在 x中的一个根。二次方程至多有两个不同的根，这与三个两两不同的相关性相矛盾。

直观地说，只有当令牌价值像 $\log b_i$ 一样增长时，比例分配才会使边际增益相等，

而这种增长永远不会饱和。若令牌价值会饱和，比例分配仍会持续向最相关的项目添加令牌，即使这些项目的令牌价值已经饱和；而这些令牌则来自相关性较低、仍能从中受益的项目。

引理 4 (分层分配)。 假设同一层级的两个项目具有不同的相关性 $\bar{r}_i \neq \bar{r}_j$、共享相同的参数 τ，并获得相同的令牌数量 b，且满足 $0 < b < \min(b_i^{\max}, b_j^{\max})$。则该分配不会最大化公式 (16)。

证明。 最优分配要求两个项目具有相等的边际增益，$\bar{r}_i\tau/(b+\tau)^2 = \bar{r}_j\tau/(b+\tau)^2$,这仅在 $\bar{r}_i = \bar{r}_j$ 时成立。

直观地说，两个项目中相关性更高的项目从下一个令牌中获得的增益仍然更大，因此将一个令牌移给它会提高目标函数。

在这两种情况下，公式（20）中的注水分配都是唯一最大化解，因此在相同预算内能够达到严格更高的目标函数值。

B. 实现细节

B.1. Lembas 的设置

表 4 列出了 Lembas 的设置。SigLIP 2 对查询和候选帧进行嵌入，BGE 对查询和候选字幕进行嵌入。在每种模态内，我们将查询与候选项目之间的余弦相似度标准化为均值为零、方差为一，并使用温度 r_i（q）通过 softmax 将其转换为检索相关性 T。令牌价值参数 τ_{frame} 和τ_{text} 决定项目的令牌价值达到饱和的速度，并通过最小二乘法，根据正文第 §3.2 节中的验证测量结果离线拟合得到。字幕权重 β用于缩放字幕相对于帧的相关性，冗余惩罚 η 用于缩放项目因相似项目而获得的折扣。迭代注水运行 L 次迭代，并在每次迭代中通过二分法找到 λ。在多轮对话中，注意力权重 ω 控制注意力对缓存项目相关性的贡献。令牌预算 B 决定每项实验的计算量，我们在每个结果中以万亿次浮点运算（TFLOPs）报告该数值。

设置	值
帧相关性编码器	SigLIP 2（so400m-patch14-384） [46]
字幕相关性编码器	BGE（base-en-v1.5） [53]
相关性温度 T	0.63
令牌价值参数 τ_{frame}	38
令牌价值参数 τ_{text}	13.6
字幕权重 β	0.5
冗余惩罚 η	2.67×10^{-5}
迭代次数 L	20
每次迭代的二分步骤数	60
注意力权重 ω (多轮)	0.5

表4. Lembas 的设置。

提示词：选择题作答

⟨4.0 秒⟩ [帧] ⟨12.0 秒⟩ [帧]...

视频字幕： *[字幕]*

根据视频，从以下多项选择题中选出最佳答案——题。回答时仅填写正确选项的字母

。

问题： *[问题]*

选项:
A. *[选项]*
B. *[选项]*
...

最佳答案是:

图9. 所有方法共用的回答提示词。灰色斜体文本标示每个问题的可变输入。

B.2. 回答提示词

所有方法都使用图 9中的提示词回答选择题。每帧前都会标注其时间戳；在 Video-MME 上，所选字幕位于问题之前。预测结果为视觉语言模型赋予最高概率的选项字母。

B.3. 基线实现

我们在相同条件下比较所有方法，因此准确率的差异取决于它们选择了哪些证据以及如何分配令牌预算。所有方法都从相同的候选帧中进行选择，并使用相同的视觉语言模型和提示词（§B.2），且令牌预算相同。在 Video-MME 上，每个基线方法还会接收与其所选帧对齐的字幕，且字幕也计入该预算。所有使用图文编码器评估查询与帧之间相关性的方法，都采用与 Lembas 相同的 SigLIP 2 相似度，因此没有方法会因使用更强的编码器而占优。QuoTA 保留其自身的评分器——作为方法本身组成部分的视觉语言模型评审器。对于多次调用视觉语言模型的方法，每次调用都会计入其计算量。除此之外，每个基线方法均采用其作者推荐的设置，具体实现如下。

- **均匀采样**：以相同间隔选取帧，每帧的令牌数量相同。
- **Top-K**：如正文所述。
- **FOCUS** [65]：使用已发布的代码，并以 SigLIP 2 的相似度替代其 BLIP 分数。
- **Q-Frame** [62]：已发布代码，其中使用 SigLIP 2 相似度替代其 Long-CLIP 分数，并将其三个分辨率层级缩放以填满令牌预算。

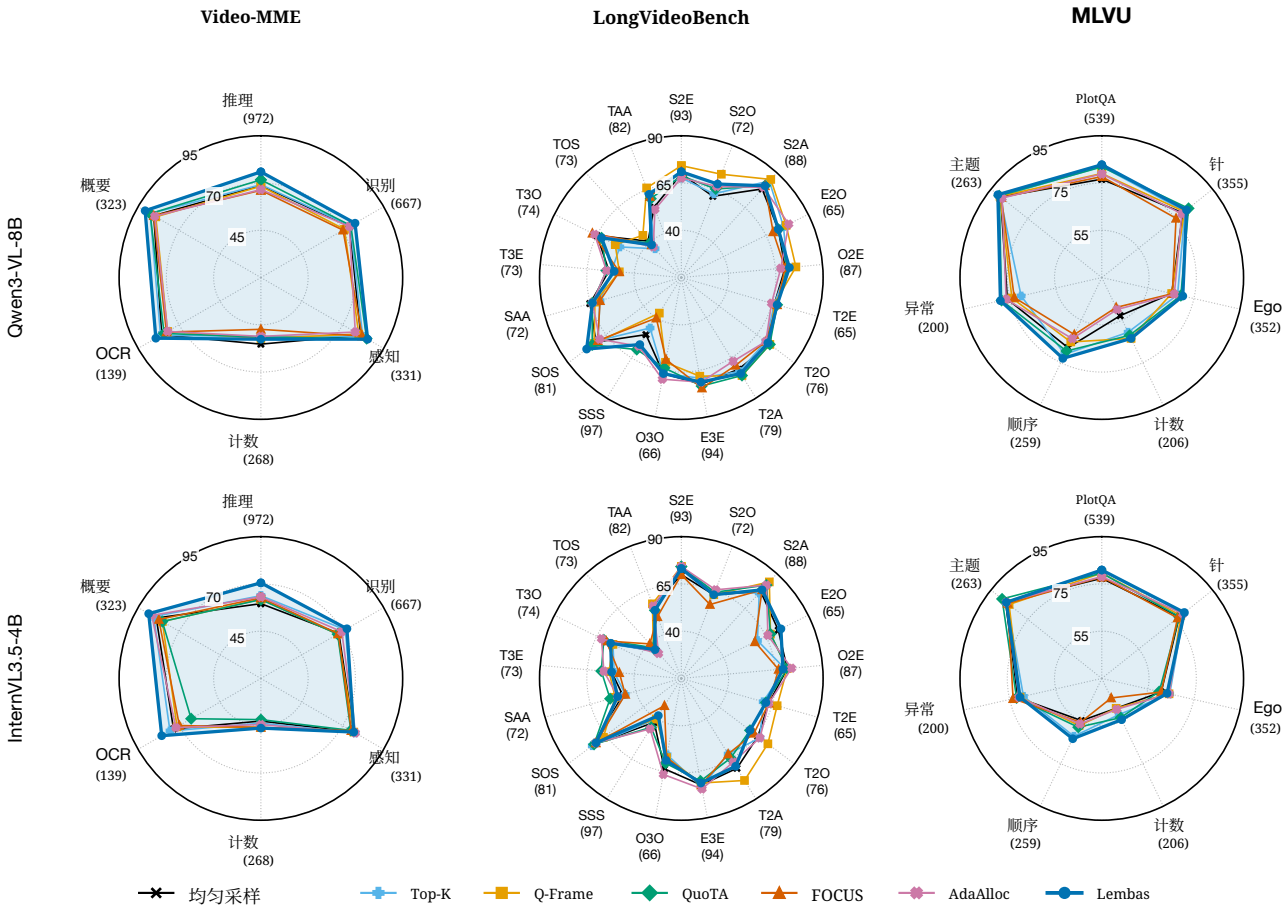

图 10。**Lembas 在比任何基线方法都更多的问题类别中取得了最高准确率**。各问题类别的准确率（%），括号内为问题数量。

• **QuoTA** [33]：已发布代码，其中包含其默认提示词和 Qwen2-VL-2B 评审模型。

•**AdaAlloc** [1]：根据其论文重新实现，并使用 SigLIP 2 相似度替代其 PE-Core 分数。由于它针对每个问题调用视觉语言模型三次，因此我们调整每次调用的令牌预算，使其总计算量与其他方法一致。

C. 补充结果

C.1. 类别准确率

图 10 按问题类别细分了正文表 1 中的准确率，其中 Video- MME 的 12 种任务类型按照 Q-Frame 的方式归为六个类别 [62]。在这三个基准测试中，Lembas 在比任何基线方法更多的类别中取得了最高准确率。在 Video- MME 上，使用两种视觉语言模型时，Lembas 都在六个类别中的五个类别中领先；其最大提升出现在推理问题上（相较于使用 Qwen3-VL-8B 和 InternVL3.5-4B 时的最佳基线，分别提升 4.2 和 6.8 个百分点）。在 MLVU 上，Lembas 使用 Qwen3-VL-8B 时在七个类别中的六个类别中领先（其中一个类别与 Q-Frame 并列），使用 InternVL3.5-4B 时则在四个类别中领先。因此，Lembas 的提升在各个问题类别中都成立，并非来自某一个单独类别。

C.2. 求解器效率

我们在包含 30 至 3,600 个候选帧的合成视频上，将迭代注水法与三种现成求解器进行比较。Gurobi [17] 和 SCIP [3] 是能够将离散分配问题求至最优的整数求解器。使用 Clarabel 求解器的 CVXPY [8] [15] 求解与迭代注水法相同的连续松弛问题，因此其目标值是整数最优值的上界。如表 5所示，迭代注水法的速度比 CVXPY 快 1.3–7× 倍，比 Gurobi 快 4–23× 倍。SCIP 在超过 420 帧时无法在 600 秒内完成，表 5 报告了它在此之前找到的最佳解。迭代注水法得出的结果也与 Gurobi 的整数最优值精确到小数点后三位，并且与上界的差距不超过 0.001。

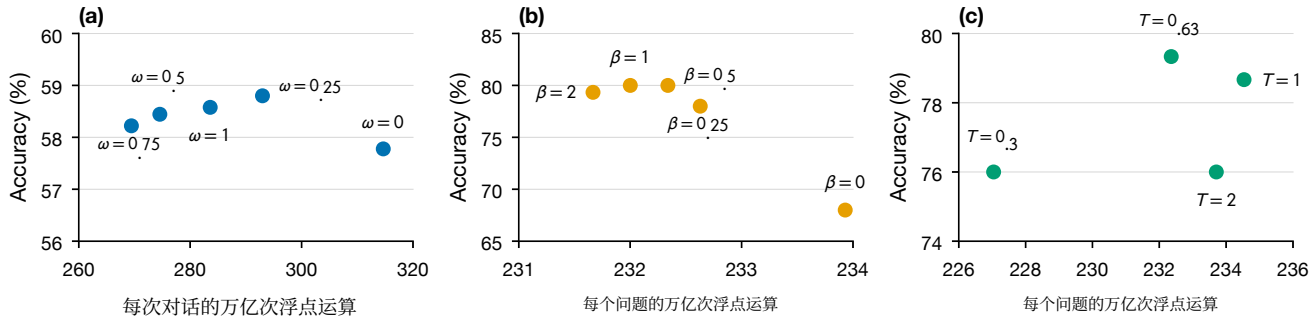

图 11。**不同超参数下 Lembas 的性能**。在以下设置变化时，准确率与计算量的关系：(a) 在 MT-Video-Bench 上改变注意力权重 ω；(b) 在 Video-MME 上改变字幕权重 β；(c) 在 Video-MME 上改变相关性温度 T。

帧	30	60	120	240	420	1,800	3,600
求解时间（秒）							
Gurobi	0.038	0.031	0.089	0.199	0.557	3.013	5.029
SCIP	1.943	3.989	9.655	23.927	51.662	600+	600+
CVXPY	0.010	0.013	0.039	0.040	0.080	0.763	1.794
Ours	**0.006**	**0.007**	**0.009**	**0.031**	**0.025**	**0.132**	**0.256**
目标函数							
Gurobi	0.698	0.679	0.595	0.607	0.635	0.456	0.461
SCIP	0.698	0.679	0.595	0.607	0.635	0.000	0.083
CVXPY	0.698	0.679	0.596	0.607	0.636	0.457	0.461
Ours	0.698	0.679	0.595	0.607	0.635	0.456	0.461

表 5。**Lembas 能够快速且近乎最优地解决分配问题**。迭代注水（我们的方法）与三种现成求解器的求解时间和目标值。

C.3. 敏感性分析

我们研究 Lembas 的准确率和计算量如何随三个超参数变化，分别是注意力权重 ω、字幕权重 β和相关性温度 T。图 11 分别改变每个超参数，同时固定其他参数。每次扫描均在 500 个问题的子集上进行。对于注意力权重（图 11a），每个 $\omega > 0$ 在准确率和计算量方面都优于仅使用检索相关性（$\omega = 0$），且 $\omega > 0$的准确率差异不到 1 个百分点。对于字幕权重（图 11b），$\beta > 0$的准确率彼此相差不超过 2 个百分点。当 $\beta = 0$时，不会为字幕分配词元，提示词中也完全不含字幕，因此准确率会下降 12 个百分点。这种设置不同于正文表 2 中“无文本分配”一行；在后者中，视觉语言模型仍会收到与选定帧对齐的字幕，这与基线方法相同。对于相关性温度（图 11c），默认值 $T = 0.63$下的准确率最高，在 $T = 1$下仍保持在 1 个百分点以内；当 T 明显偏小或偏大时，准确率会下降约 3 个百分点。因此，Lembas 在较宽范围内对 ω 和 β 具有鲁棒性，而对接近默认值的 T 也具有鲁棒性。